개발환경보고서

|  |  |  |  |
| --- | --- | --- | --- |
| 환경 | 개발환경 OS | **window** | |
| DB | **Mysql** | |
| 개발언어 | **Python3** | -Python 2와 하위호환성이 없음.  -교차 플랫폼 언어  -C, C++, Java 같은 언어와 쉽게 통합 |
| 라이브러리 | **Beautiful Soup** | -HTML이나 XML에서 데이터를 추출하는 Python Library.  -웹문서의 구조를 찾아내는 파서를 이용해 찾고자 하는 데이터의 위치를 찾아 내어 값을 추출. |
| **Selenium** | -Web앱을 테스트하는데 이용하는 프레임워크.  -‘webdriver’ 라는 API를 통해 운영체제에 설치된 Chrome등의 브라우저를 제어 가능.  -브라우저를 직접 동작시키기 때문에 JavaScript를 이용해 비동기적으로 혹은 뒤늦게 불러와지는 컨텐츠들을 가져올 수 있음.  -JS로 렌더링이 완료된 후의 DOM 결과물에 접근이 가능.  -자체적인 웹브라우저가 들어있지 않으므로, 다른 브라우저(파이어폭스, 익스플로러, 크롬…)가 있어야 사용할 수 있음.  \* 웹 브라우저 라이브러리 후보 2개  **1) 팬텀 js**  - 스트립트가 백그라운드에서 조용히 실행 되는 것을 원할 때 사용  - 인터페이스가 없는 headless 브라우저.  - 웹사이트를 메모리에 물러오고 페이지의 자바 스크립트를 실행하지마, 그래픽은 렌더링하지 않음.  - 팬텀 또한 파이썬 라이브러리가 아니므로 직접 내려 받아야 함.  **2) 구글 크롬 웹 드라이버**  - 인터페이스가 있으므로, 브라우저 화면이 앞에 있기 때문에 디버그 하기가 편함.  - 코드를 짠 예상대로 흐름이 흘러가지 않을 때, 코드 실행을 잠시 멈추고 직접 조작해 볼 수 있음. |
| **Tweepy** | -Twitter에게 API를 요청하고 정보를 얻는 모듈 (단, 7일 내의 자료만 얻을 수 있다는 단점이 있음) |
| **Pandas** | -엑셀, 텍스트 파일, SQL 데이터 베이스 등의 데이터를 분석하기 위한 라이브러리.  -행과 열로 이루어진 데이터 객체를 만들어 다룰 수 있어 보다 안정적으로 대용량 데이터들을 처리 가능한 도구 |
| **Numpy** | -행렬이나 일반적으로 대규모 다차원 배열을 쉽게 처리 할 수 있도록 지원하는 라이브러리.  -데이터 구조 외에도 수치 계산을 위해 효율적으로 구현된 기능 제공 |
| **CSV** | -CSV파일을 완벽하게 읽고 쓸 수 있는 라이브러리.  -파일을 문자열 형식으로 읽은 후 StringIO 객체로 바꿔서 파일처럼 다뤄서 이용 |
| **Facebook Graph API** | -데이터를 가져오거나 게시하는 HTTP기반의 API.  -Node와 Edge에 대하여 GET요청을 통해 데이터를 가져옴. |
| 텍스트에디터 | **파이참** | -패키지 설치가 쉽고 코드를 실행하는 방법이 간편하며, python파일과 함수관리가 쉬움. |
| DB정보 | DBMS | **MySQL** | -다른 제품들과 퍼포먼스의 큰 차이는 없지만 비용면에서 저렴하고 많은 OS 플랫폼에서 사용할 수 있음 |
| DB library | **pymysql** | -파이썬 프로그램에서 mySQL 서버와 통신할 수 있게 해주는 파이썬용 데이터베이스 커넥터 |

|  |  |
| --- | --- |
| 기타정보 | 1. CSV(comma-separated values) : 필드를 쉼표(,)로 구분한 텍스트 데이터 및 텍스트 파일이다.    2. Pandas 는 빅데이터 분석을 위한 파이선 라이브러리이다.    - read\_csv(파일경로명)을 통해 csv 파일을 파이썬의 dataframe 형태로 바꿔준다.  - csv 파일의 인덱스별로 column이 만들어지고 그 안에 값이 저장된다.  3. csv 파일이 아니더라도 delimiter를 무엇으로 할지 정한다면 다른 형태의 파일도 변경가능하다.  4. 만약 csv 파일에 인덱스가 없을 경우  1)    이처럼 header에 None을 넘겨주면 자동으로 인덱스가 생성된다.  2)  이처럼 read\_csv의 파라미터로 인덱스 명을 정해줄 수 있다.  5. to\_csv()를 통해 read\_csv()의 반대 작업이 가능하다.  - to\_csv()는 파이쎤의 dataframe 정보를 csv 형태로 변환해준다. |

<데이터 전달양식>

**플랫폼**

**웹사이트----4)**

1. 전체 개발순서

1) 크롤링을 이용하여 데이터수집

2) 수집한 데이터를 가지고 통계적 분석 모델 또는 머신러닝

3) 분석모델을 다른곳에 적용할 수 있도록 **json파일**을 이용하여 데이터 저장

4) 수집한 결과를 json파일에 적용, **rest api**를 호출하여 이를 보여주는 페이지를 이용하여 시각화.

- Rest api

: 서버를 오픈해두고, 다른 곳에서 서버에 request를 보내면 데이터를 처리해서 그 결과를 보내줌.

Response 시 json형식으로 미리 정의해둔것에 맞춰서 돌림.

5) 수집한 결과를 json파일에 적용, **rest api**로 오픈하여 다른 곳에서 사용할 수 있도록 제공